On the Generalization of the C-Bound to Structured Output Ensemble Methods

François Laviolette[1] Emilie Morvant[2] Liva Ralaivola[3] Jean-Francis Roy[1]
[1] Département d'informatique et de génie logiciel, Université Laval, Québec, Canada
[2] Laboratoire Hubert Curien, Université Jean Monnet, UMR CNRS 5516, Saint-Etienne, France
[3] Aix-Marseille Univ., LIF-QARMA, CNRS, UMR 7279, F-13013, Marseille, France

June 15, 2015

Abstract

This paper generalizes an important result from the PAC-Bayesian literature for binary classification to the case of ensemble methods for structured outputs. We prove a generic version of the $\mathcal{C}$-bound, an upper bound over the risk of models expressed as a weighted majority vote that is based on the first and second statistical moments of the vote's margin. This bound may advantageously (i) be applied on more complex outputs such as multiclass labels and multilabel, and (ii) allow to consider margin relaxations. These results open the way to develop new ensemble methods for structured output prediction with PAC-Bayesian guarantees.

1 Introduction

It is well-known that learning predictive models capable of dealing with outputs that are richer than binary outputs (*e.g.,* multiclass or multilabel) and for which theoretical guarantees exist is still a realm of intensive investigations. From a practical standpoint, a lot of relaxations for learning with complex outputs have been devised. A common approach consists in decomposing the output space into "simpler" spaces so that the learning problem at hand can be reduced to a few easier (*i.e.,* binary) learning tasks. For instance, this is the idea spurred by the Error-Correcting Output Codes (Dietterich & Bakiri, 1995) that makes possible to reduce multiclass or multilabel problems into binary classification tasks,*e.g.*, (Allwein et al., 2001; Mroueh et al., 2012; Read et al., 2011; Tsoumakas & Vlahavas, 2007; Zhang & Schneider, 2012). In our work, we study the problem of complex output prediction by focusing on prediction functions that take the form of a weighted majority vote over a set of complex output classifiers (or voters). Recall that *ensemble methods* can all be seen as majority vote learning procedures (Dietterich, 2000; Re & Valentini, 2012). Methods such as Bagging (Breiman, 1996), Boosting (Schapire & Singer, 1999) and Random Forests (Breiman, 2001) are representative voting methods. Cortes et al. (2014) have proposed various ensemble methods for the structured output prediction framework.

Note also that majority votes are also central to the Bayesian approach (Gelman et al., 2004) with the notion of Bayesian model averaging (Domingos, 2000; Haussler et al., 1994) and most of kernel-based predictors, such as the Support Vector Machines (Boser et al., 1992; Cortes & Vapnik, 1995) may be viewed as weighted majority votes as well: for binary classification, where the predicted class for some input $\mathbf{x}$ is computed as the sign of $\sum_i \alpha_i \, \mathbf{y}_i \, k(\mathbf{x}_i, \mathbf{x})$, each voter is simply given by $\mathbf{y}_i k(\mathbf{x}_i, \cdot)$.

From a theoretical standpoint, as far as binary classification is concerned, the notion of *margin* is often the crux to establish the generalization ability of a majority vote predictor. For instance, still considering the binary case, the margin of a majority vote realized on an example is defined as the difference between the total weight of the voters that predicted the correct class minus the total weight given to the incorrect one. In the PAC-Bayesian analysis, which is our working setup, one way to provide generalization bounds for a majority vote classifier is to relate it to a stochastic classifier, the *Gibbs* classifier, whose risk is the weighted risk of the individual voters involved in the majority vote. Up to a linear transformation, the Gibbs risk is equivalent to the first statistical moment of the margin (Laviolette et al., 2011). This PAC-Bayesian analysis can be very accurate when the Gibbs risk is low, as in the situation where the voters having large weights are performing well as done by Germain et al. (2009), Langford & Shawe-Taylor (2002), together with McAllester (2009) for the structured output framework. However, for ensemble methods, it is not unusual to be in the situation where, on the one hand, the voters achieve performances

only slightly above the chance level—which makes it impossible to find weights that induce a small Gibbs risk—and, on the other hand, the risk of the majority vote itself is very low. Hence, to better capture the accuracy of a majority vote in a PAC-Bayesian fashion, it is required to consider more than the Gibbs risk, *i.e.*, more than only the first statistical moment of the margin. This idea, which has been studied in the context of ensemble methods by (Blanchard, 2004; Breiman, 2001), has been revisited as the $\mathcal{C}$-bound by Lacasse et al. (2007) in the PAC-Bayesian setting. This bound sheds light on an essential feature of weighted majority votes: how good the voters individually are is just as important as how correlated their predictions are. This has inspired a new ensemble method for binary classification with PAC-Bayesian generalization guarantees named MinCq (Laviolette et al., 2011), whose performances are state-of-the-art. In the multiclass setting, there exists one PAC-Bayesian bound, but it is based on the confusion matrix of the Gibbs classifier (Morvant et al., 2012). Kuznetsov et al. (2014) have recently proposed an improved Rademacher bound for multiclass prediction that is based on the notion of the multiclass margin of Breiman (2001) (Definition 1 in the present paper). However, as for the binary case, these bounds suffer from the same lack of tightness when the voters of the majority vote perform poorly.

Here, we intend to generalize the $\mathcal{C}$-bound to more complex situations. We first propose a formulation of the $\mathcal{C}$-bound for ensemble methods for complex output settings, that makes it possible for all the binary classification-based results of Lacasse et al. (2007) and Laviolette et al. (2011) to be generalized. Since for complex output prediction the usual margin relies on the maximal deviation between the total weight given to the true class minus the maximal total weight given to the "runner-up" incorrect one, we base our theory on a general notion of margin allowing us to consider extensions of the usual margin. Moreover, similarly as for binary classification (Lacasse et al., 2007; Laviolette et al., 2011), we derive a PAC-Bayesian generalization bound and show how we can estimate such $\mathcal{C}$-bounds from a sample. In light of this general theoretical result, we propose two specializations suitable for multiclass classification with ensemble methods based on the true margin and on a relaxation that we call ω-margin. We highlight the behavior of these $\mathcal{C}$-bounds through an empirical study. Finally, we instantiate it to multilabel prediction problems. We see these theoretical results as a first step towards the design of new and well-founded learning algorithms for complex outputs.

This paper is organized as follows. Section 2 recalls the binary $\mathcal{C}$-bound, that is generalized to a more general setting in Section 3. We then specialize this bound to multiclass prediction in Section 4, and to multilable prediction in Section 5. We conclude in Section 6.

2 Ensemble Methods in Binary Classification

For binary classification with majority vote-based ensemble methods, we often consider an arbitrary input space $\mathbf{X}$, an output space of two classes $\mathbf{Y} = \{-1, +1\}$, and a set $\boldsymbol{\mathcal{H}} \subseteq \{\mathbf{h} : \mathbf{X} \to [-1, +1]\}$ of *voters*. A voter can return any value in $[-1, +1]$, interpretable as a level of confidence of the voter into its option which is $+1$ if the output is positive and -1 otherwise. A voter that always outputs values in $\{-1, +1\}$ is a *(binary) classifier*. The *binary ρ-weighted majority vote* $\mathbf{B}_\rho(\cdot)$ is the classifier returning either of the two options that has obtained the larger weight in the vote, *i.e.*:

$$\forall \mathbf{x} \in \mathbf{X}, \quad \mathbf{B}_\rho(\mathbf{x}) \;=\; \underset{\mathbf{y}\in\{-1,+1\}}{\operatorname{argmax}} \; \underset{\mathbf{h}\sim\rho}{\mathbf{E}} \Big(\big|\mathbf{h}(\mathbf{x})\big| \, \mathbf{I}\Big[\operatorname{sign}(\mathbf{h}(\mathbf{x})) = \mathbf{y} \Big] \Big) \;=\; \operatorname{sign}\left[\underset{\mathbf{h}\sim\rho}{\mathbf{E}} \mathbf{h}(\mathbf{x}) \right] ,$$

where $\mathbf{I}[a] = 1$ if predicate a is true and 0 otherwise.
Given a training set S of observed data in which each example $(\mathbf{x}, \mathbf{y}) \in S$ is drawn *i.i.d.* from a fixed yet unknown probability distribution D on $\mathbf{X} \times \{-1, +1\}$, the learner aims at finding a weigthing distribution ρ over $\mathcal{H}$ that induces a low-error majority vote; in other words, minimizing the *true risk* of the ρ-weighted majority vote $\mathbf{R}_D(\mathbf{B}_\rho) = \mathbf{E}_{(\mathbf{x},\mathbf{y})\sim D} \, \mathbf{I}\left[\mathbf{B}_\rho(\mathbf{x}) \neq \mathbf{y}\right]$ under the 0-1-loss is aimed for. One way towards this goal is to implement the Empirical Risk Minimization (ERM) principle, that is to minimize the *empirical risk* of the majority vote $\mathbf{R}_S(\mathbf{B}_\rho)$ estimated on the sample S. Unfortunately, a well-known issue to learn such weights is that the direct minimization of $\mathbf{R}_S(\mathbf{B}_\rho)$ is an $\mathcal{NP}$-hard problem. To overcome this, we may use relaxations of the risk, look for estimators or bounds of the true risk that are simultaneously valid for all possible distributions ρ on $\boldsymbol{\mathcal{H}}$, and try to minimize them. In the PAC-Bayesian theory (McAllester, 1999), such an estimator is given by the *Gibbs risk* $\mathbf{R}(\mathbf{G}_\rho)$ of a ρ-weighted majority vote which is simply the ρ-average risk of the voters. Indeed, it is well known that the risk of the majority vote is bounded by twice its Gibbs risk:

$$\mathbf{R}_D(\mathbf{B}_\rho) \;\leq\; 2\,\mathbf{R}_D(\mathbf{G}_\rho) \;=\; 2\,\mathbf{E}_{\mathbf{h}\sim\rho}\,\mathbf{R}_D(\mathbf{h})\,. \tag{1}$$

With this relation, PAC-Bayesian theory indirectly gives generalization bounds for ρ-weighted majority votes. Unfortunately, even if they tightly bound the true risk $\mathbf{R}_D(\mathbf{G}_\rho)$ in terms of its empirical counterpart $\mathbf{R}_S(\mathbf{G}_\rho)$, this tightness might not carry over to the bound on the majority vote.

Note that even if there exist situations for which Inequality (1) is an equality, ensemble methods (especially when the voters are weak) build on the idea that the risk of the majority vote might be way below the average of its voters' risk. Indeed, it is well known that voting can dramatically improve performances when the "community" of voters tends to compensate the individual errors. The "classical" PAC-Bayesian framework of McAllester (1999) does not allow to evaluate whether or not this compensation occurs. To overcome this problem, Lacasse et al. (2007) proposed not only to take into account the mean of the errors of the associated Gibbs $\mathbf{R}_D(\mathbf{G}_\rho)$, but also its variance. They proposed a new bound, called the *$\mathcal{C}$-bound*, that replaces the loose factor of 2 in Inequality (1). They also extended the PAC-Bayesian theory in such a way that both the mean and the variance of the Gibbs classifier can be estimated from the training data simultaneously for all ρ's. Laviolette et al. (2011) have reformulated this approach in terms of the first and second statistical moment of the *margin* realized by the ρ-weighted majority vote, which pertains to known results in non-PAC-Bayesian frameworks, *e.g.,* (Blanchard, 2004; Breiman, 2001); the margin of a ρ-weighted majority vote on an example $(\mathbf{x}, \mathbf{y}) \in \mathbf{X} \times \mathbf{Y}$ being:

$$\mathbf{M}_\rho(\mathbf{x}, \mathbf{y}) \;=\; \mathbf{y} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(\mathbf{x})\,.$$

In terms of this margin, the $\mathcal{C}$-bound is defined as follows.

Theorem 1 ($\mathcal{C}$-bound of Laviolette et al. (2011))**.** *For every distribution ρ on a set of voters $\mathcal{H}$, and for every distribution D on $\mathbf{X} \times \mathbf{Y}$, if $\mathbf{E}_{(\mathbf{x},\mathbf{y})}\, \mathbf{y}\, \mathbf{E}_{\mathbf{h}\sim\rho}\, \mathbf{h}(\mathbf{x}) > 0$, then we have:*

$$\mathbf{R}_D(\mathbf{B}_\rho) \;\leq\; 1 - \frac{\left(\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim D} \mathbf{y} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(\mathbf{x})\right)^2}{\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim D} \left(\mathbf{y} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(\mathbf{x})\right)^2}\,.$$

Proof. First, note that

$$\mathbf{R}_D(\mathbf{B}_\rho) = \mathop{\mathbf{Pr}}_{(\mathbf{x},y)\sim D} \left(\mathbf{M}_\rho(\mathbf{x}, y) \leq 0\right)\,,$$

then to upper-bound $\mathbf{R}_D(\mathbf{B}_\rho)$ it suffices to upper-bound $\mathbf{Pr}_{(\mathbf{x},y)\sim D}\left(\mathbf{M}_\rho(\mathbf{x}, y) \leq 0\right)$. Making use of the Cantelli-Chebyshev inequality that states that for any random variable Z,

$$\forall a > 0,\ \mathbf{Pr}\left(Z \leq \mathbf{E}\,[Z] - a\right) \leq \frac{\mathbf{Var}\, Z}{\mathbf{Var}\, Z + a^2}\,,$$

we get the desired result if $Z = \mathbf{M}_\rho(\mathbf{x}, \mathbf{y})$, with $a = \mathbf{E}_{(\mathbf{x},\mathbf{y})\sim D}\, \mathbf{M}_\rho(\mathbf{x}, \mathbf{y})$ combined with the definition of the variance. The constraint $\mathbf{E}_{(\mathbf{x},\mathbf{y})}\, \mathbf{y}\, \mathbf{E}_{\mathbf{h}\sim\rho}\, \mathbf{h}(\mathbf{x}) > 0$ comes from the fact that this inequality is valid for $a > 0$. □

The $\mathcal{C}$-bound involves both the ρ-weighted majority vote confidence via $\mathbf{E}_{(\mathbf{x},\mathbf{y})}(\mathbf{y}\, \mathbf{E}_{\mathbf{h}\sim\rho}\, \mathbf{h}(\mathbf{x}))$ and the correlation between the voters via $\mathbf{E}_{(\mathbf{x},\mathbf{y})}(\mathbf{y}\, \mathbf{E}_{\mathbf{h}\sim\rho}\, h(\mathbf{x}))^2$. It is known to be very precise. Minimizing its empirical counterpart then appears as a natural solution for learning a distribution ρ leading to a well-performing binary ρ-weighted majority vote. Moreover, this strategy is justified by a PAC-Bayesian generalization bound over the $\mathcal{C}$-bound (similar to Theorem 3 of this paper but restricted to the case where $\mathbf{Y} = \{-1, +1\}$), and has given the MinCq algorithm (Laviolette et al., 2011).

As announced earlier, we here intend to generalize the $\mathcal{C}$-bound theory to more complex outputs than binary outputs. Our contributions first consist in generalizing—in Section 3—this important result to a broader ensemble method setting, along with PAC-Bayesian generalization bounds.

3 A General Setting for Majority Votes over a Set of Complex Output Voters

In this section, we propose a general setting in which one can consider classification under ρ-weighted majority votes. We present a general definition of the margin and propose a $\mathcal{C}$-bound designed for majority vote-based ensemble methods when we want to combine complex output classifiers. We also discuss how to estimate this bound from a set S of m examples drawn *i.i.d.* from D. To do so, we derive a PAC-Bayesian theorem that bounds the *true* risk of the ρ-weighted majority vote $\mathbf{R}_D(\mathbf{B}_\rho)$ by using the empirical estimation of our new $\mathcal{C}$-bound on the sample S.

3.1 A General $\mathcal{C}$-bound for Complex Output Prediction

Given some input space $\mathbf{X}$ and a *finite* output space $\mathbf{Y}$, we suppose that there exists a feature map $Y : \mathbf{Y} \to H_{\mathbf{Y}}$, where $H_{\mathbf{Y}}$ is a vector space such as a Hilbert space. For the sake of clarity, we suppose that all the vectors $Y(\mathbf{Y})$ are unit-norm vectors; most of the following results remain true without this assumption but have to be stated in a more complicated form. Let $\operatorname{Im}\mathbf{Y}$ be the image of $\mathbf{Y}$ under $Y(\cdot)$, and $\operatorname{conv}(\operatorname{Im}\mathbf{Y})$ ($\subseteq H_{\mathbf{Y}}$) its convex hull. We consider a (non-necessarily finite) set of *voters* $\boldsymbol{\mathcal{H}} \subseteq \{\mathbf{h} : \mathbf{X} \to \operatorname{conv}(\operatorname{Im}\mathbf{Y})\}$. A voter that always outputs values in $\operatorname{Im}\mathbf{Y}$ is called a classifier. For every probability distribution ρ on $\boldsymbol{\mathcal{H}}$, we define the *ρ-weighted majority vote classifier* $\mathbf{B}_\rho$ such that:

$$\forall \mathbf{x} \in \mathbf{X}, \quad \mathbf{B}_\rho(\mathbf{x}) = \operatorname*{argmin}_{\mathbf{c}\in\mathbf{Y}} \left\| Y(\mathbf{c}) - \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(\mathbf{x}) \right\|^2 = \operatorname*{argmax}_{\mathbf{c}\in\mathbf{Y}} \left\langle Y(\mathbf{c}) \,,\, \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(\mathbf{x}) - \frac{1}{2} Y(\mathbf{c}) \right\rangle . \tag{2}$$

As in the binary classification case, the learning objective is to find a distribution ρ that minimizes the *true risk* $\mathbf{R}_D(\mathbf{B}_\rho)$ of the ρ-weighted majority vote given by

$$\mathbf{R}_D(\mathbf{B}_\rho) \;=\; \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim D} \mathbf{I}\left[\mathbf{B}_\rho(\mathbf{x}) \neq \mathbf{y}\right] .$$

Inspired by the margin definition of Breiman (2001), we propose the following generalization of the binary margin, which measures the confidence of a prediction as the deviation between the voting weights received by the correct class and the largest voting weight received by any incorrect class.

Definition 1. *For any example $(\mathbf{x}, \mathbf{y})$ and any distribution ρ on a set of voters $\boldsymbol{\mathcal{H}}$, we define the* margin $\mathbf{M}_\rho(\mathbf{x}, \mathbf{y})$ of the ρ-weighted majority vote *on* $(\mathbf{x}, \mathbf{y})$ *as:*

$$\begin{aligned} \mathbf{M}_\rho(\mathbf{x}, \mathbf{y}) &= \left\langle Y(\mathbf{y}) \,,\, \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(\mathbf{x}) - \tfrac{1}{2} Y(\mathbf{y}) \right\rangle - \max_{\substack{\mathbf{c}\in\mathbf{Y} \\ \mathbf{c}\neq\mathbf{y}}} \left\langle Y(\mathbf{c}) \,,\, \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(\mathbf{x}) - \tfrac{1}{2} Y(\mathbf{c}) \right\rangle \\ &= \left\langle \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(\mathbf{x}) \,,\, Y(\mathbf{y}) \right\rangle - \max_{\substack{\mathbf{c}\in\mathbf{Y} \\ \mathbf{c}\neq\mathbf{y}}} \left\langle \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(\mathbf{x}) \,,\, Y(\mathbf{c}) \right\rangle . \end{aligned} \tag{3}$$

The second form of the margin readily comes from simple computations combined with our assumption that $\left\|Y(\mathbf{y})\right\| = 1$, $\forall \mathbf{y} \in \mathbf{Y}$.

With this definition at hand, it is obvious that the ρ-weighted majority vote errs on $(\mathbf{x}, \mathbf{y})$ if and only if the margin realized on $(\mathbf{x}, \mathbf{y})$ is negative. Therefore,

$$\mathbf{R}_D(\mathbf{B}_\rho) \;=\; \mathop{\mathbf{Pr}}_{(\mathbf{x},\mathbf{y})\sim D} \left(\mathbf{M}_\rho(\mathbf{x}, \mathbf{y}) \leq 0\right) . \tag{4}$$

Remark 1. *We may retrieve the binary notion of majority vote from our general framework in various ways. For example, one may take $Y : \{-1, +1\} \to \mathbb{R}$ with $Y(+1) = 1$ and $Y(-1) = -1$ considering each binary voters as a voter. Note that the definition of the margin, and, henceforth, that of the $\mathcal{C}$-bound, will differ because the binary definition of the margin is linear in $\mathbf{E}_{\mathbf{h}\sim\rho}\, \mathbf{h}(\mathbf{x})$ whereas our margin definition is quadratic in that variable. To fall back to the exact same $\mathcal{C}$-bound from our framework, the squared norm should be replaced by the norm itself in Definition 1. We however purposely choose to work with the square of the norm as it renders the calculations easier.*

Using the proof technique of Theorem 1, we arrive at a general $\mathcal{C}$-bound.

Theorem 2 (General $\mathcal{C}$-bound)**.** *For every probability distribution ρ over a set of voters $\boldsymbol{\mathcal{H}}$ from $\mathbf{X}$ to* $\operatorname{conv}(\operatorname{Im}\mathbf{Y})$*, and for every distribution D on $\mathbf{X} \times \mathbf{Y}$, if $\mathbf{E}_{(\mathbf{x},\mathbf{y})\sim D}\, \mathbf{M}_\rho(\mathbf{x}, \mathbf{y}) > 0$, then we have:*

$$\mathbf{R}_D(\mathbf{B}_\rho) \;\leq\; 1 - \frac{\left(\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim D} \mathbf{M}_\rho(\mathbf{x}, \mathbf{y}) \right)^2}{\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim D} \left(\mathbf{M}_\rho(\mathbf{x}, \mathbf{y})\right)^2} .$$

Proof. Thanks to Equation (4) the proof consists in bounding $\mathbf{Pr}_{(\mathbf{x},\mathbf{y})} \left(\mathbf{M}_\rho(\mathbf{x}, \mathbf{y}) \leq 0\right)$ with the Cantelli-Chebyshev inequality as done for Theorem 1. □

Remark 2 (On the construction of the set of voters $\mathcal{H}$). *All our results hold for both extreme cases of weak voters, as usual in ensemble methods, and that of more expressive/highly-performing voters. Typical instantiations of the former situation are encountered when making use of a kernel function $k : X \times X \to \mathbb{R}$ that induces the set of voters $\mathcal{H} = \{\, k(\mathbf{x}, \cdot)\, Y(\mathbf{y}) \,|\, (\mathbf{x}, \mathbf{y}) \in S\}$; the situation also arises when a set of structured prediction functions learned with different hyperparameters are considered; as evidenced in Section 4.2 for the multiclass setting, the weak voters may also be decision stumps or (more-or-less shallow) trees. Combining more expressive voters is a situation that may show up as a need to combine voters obtained from a primary mechanism. This is for instance the case in multiview learning (Sun, 2013) when we want to combine models learned from several data descriptions—note that, the binary $\mathcal{C}$-bound has already shown its relevance in such a situation (Morvant et al., 2014).*

3.2 A PAC-Bayesian Theorem to Estimate the General $\mathcal{C}$-bound

In this section, we briefly discuss how to estimate the previous bound from a sample S constituted by m examples drawn *i.i.d.* from D. To reach this goal, we derive a PAC-Bayesian theorem that upper-bounds the *true* risk $\mathbf{R}_D(\mathbf{B}_\rho)$ of the ρ-weighted majority vote by using the empirical estimation of the $\mathcal{C}$-bound of Theorem 2 on the sample S.

Theorem 3. *For any distribution D on $\mathbf{X} \times \mathbf{Y}$, for any set $\mathcal{H}$ of voters from $\mathbf{X}$ to $\mathrm{conv}(\mathrm{Im}\,\mathbf{Y})$, for any prior distribution π on $\mathcal{H}$ and any $\delta \in (0, 1]$, with a probability at least $1 - \delta$ over the choice of the m-sample $S \sim (D)^m$, for every posterior distribution ρ over $\mathcal{H}$, if $\mathbf{E}_{(\mathbf{x},\mathbf{y})\sim D}\, \mathbf{M}_\rho(\mathbf{x}, \mathbf{y}) > 0$, we have:*

$$\mathbf{R}_D(\mathbf{B}_\rho) \;\leq\; 1 - \frac{\left(\max\left[0, \frac{1}{m}\sum_{(\mathbf{x},\mathbf{y})\in S} \mathbf{M}_\rho(\mathbf{x},\mathbf{y}) - \sqrt{\frac{2B}{m}\left[\mathrm{KL}(\rho\|\pi) + \ln\frac{2\sqrt{m}}{\delta/2}\right]}\right]\right)^2}{\min\left[1, \frac{1}{m}\sum_{(\mathbf{x},\mathbf{y})\in S} \left(\mathbf{M}_\rho(\mathbf{x},\mathbf{y})\right)^2 + \sqrt{\frac{2B^2}{m}\left[2\mathrm{KL}(\rho||\pi) + \ln\frac{2\sqrt{m}}{\delta/2}\right]}\right]},$$

where $B \in (0, 2]$ bounds the absolute value of the margin $|\mathbf{M}_\rho(\mathbf{x}, \mathbf{y})|$ for all $(\mathbf{x}, \mathbf{y})$, and $\mathrm{KL}(\rho\|\pi) = \mathbf{E}_{\mathbf{h}\sim\rho} \ln \frac{\rho(\mathbf{h})}{\pi(\mathbf{h})}$ is the Kullback-Leibler divergence between ρ and π.

Proof. First since $\|Y(\mathbf{y})\| = 1$, $\forall \mathbf{y} \in \mathbf{Y}$, and $\mathbf{h}(\mathbf{x}) \in \mathrm{conv}(\mathrm{Im}\,\mathbf{Y})$, $\forall \mathbf{x} \in \mathbf{X}$, then $\langle \mathbf{E}_{\mathbf{h}\sim\rho}\, \mathbf{h}(\mathbf{x}),\, Y(\mathbf{c})\rangle$ takes its value between -1 and $+1$. It follows from Equation (3) that $B = 2$ is always an upper bound of $|\mathbf{M}_\rho(\mathbf{x}, \mathbf{y})|$. The bound is obtained by deriving a PAC-Bayesian lower bound on $\mathbf{E}_{(\mathbf{x},\mathbf{y})\sim D}\, \mathbf{M}_\rho(\mathbf{x}, \mathbf{y})$ and a PAC-Bayesian upper bound on $\mathbf{E}_{(\mathbf{x},\mathbf{y})\sim D} \left(\mathbf{M}_\rho(\mathbf{x}, \mathbf{y})\right)^2$. We then use the union bound argument to make these two bounds simultaneously valid, and the result follows from Theorem 2. These two bounds and their respective proof are provided in Appendix A, as Theorems 6 and 7. □

Unlike to classical PAC-Bayesian bounds and especially those provided for structured output prediction by McAllester (2009), our theorem has the advantage to directly upper-bound the risk of the ρ-weighted majority vote thanks to the $\mathcal{C}$-bound of Theorem 2. Moreover, it allows us to deal with either the general notion of margin, or margin's surrogate as illustrated in the following.

3.3 A Surrogate for the Margin

The general notion of margin can be hard to exploit in general because of its second term relying on a max. We propose to define a simpler surrogate of the margin, by replacing the second term in Equation (3) by a threshold $\omega \in [0, 1]$.

Definition 2 (The ω-margin). *For any example $(\mathbf{x}, \mathbf{y}) \in \mathbf{X} \times \mathbf{Y}$ and for any distribution ρ on $\mathcal{H}$, we define the ω-margin $\mathbf{M}_{\rho,\omega}(\mathbf{x}, \mathbf{y})$ of the ρ-weighted majority vote realized on $(\mathbf{x}, \mathbf{y})$ as*

$$\mathbf{M}_{\rho,\omega}(\mathbf{x}, \mathbf{y}) \;=\; \left\langle \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}(\mathbf{x})\,,\, \mathbf{Y}(\mathbf{y}) \right\rangle - \omega\,.$$

Trivially, the ω-margin always upper-bounds the margin when $\omega = 0$. Moreover, since $\forall Y(\mathbf{y}) \in \mathrm{Im}\,\mathbf{Y}, \|Y(\mathbf{y})\| = 1$, and $\forall \mathbf{x} \in \mathbf{X}$, $\mathbf{E}_{\mathbf{h}\sim\rho}\, \mathbf{h}(\mathbf{x}) \in \mathrm{conv}(\mathrm{Im}\,\mathbf{Y})$, then the ω-margin always lower-bounds the margin when $\omega = 1$. We

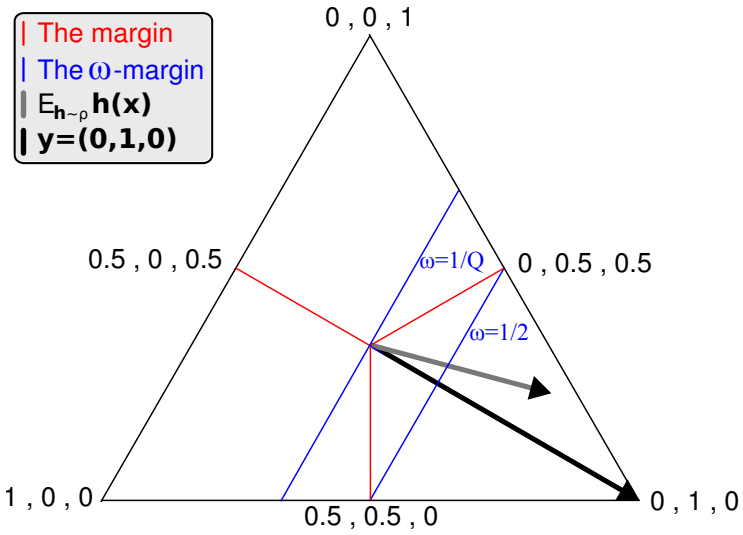

Figure 1: Representation of the multiclass margins and the vote applied on $(\mathbf{x}, \mathbf{y})$ in the barycentric coordinate system defined by $\text{conv}(\text{Im}\,\mathbf{Y})$ when $\mathbf{Y} = \{1, 2, 3\}$ and the true $\mathbf{y}$ is 2, *i.e.*, $(0, 1, 0)^\top$. We have $Y(1) = (1, 0, 0)^\top$, $Y(2) = (0, 1, 0)^\top$, and $Y(3) = (0, 0, 1)^\top$. Each line is the decision boundary of a margin: the hyperplane where lies each example with a margin equals to 0. A vote correctly classifies an example if it lies on the same side of the hyperplane than the correct class.

will see that in the multiclass setting it is also the case for $\omega = \frac{1}{2}$. When the ω-margin lower-bounds the margin, we can replace it in the $\mathcal{C}$-bound in the following way:

$$\mathcal{C}(\mathbf{M}_{\rho,\omega}) \;=\; 1 - \frac{\left(\underset{(\mathbf{x},\mathbf{y})\sim D}{\mathbf{E}} \mathbf{M}_{\rho,\omega}(\mathbf{x},\mathbf{y})\right)^2}{\underset{(\mathbf{x},\mathbf{y})\sim D}{\mathbf{E}} \left(\mathbf{M}_{\rho,\omega}(\mathbf{x},\mathbf{y})\right)^2}\,. \tag{5}$$

Indeed, in this situation we have:

$$\mathbf{R}_D(\mathbf{B}_\rho) \;=\; \underset{(\mathbf{x},\mathbf{y})\sim D}{\Pr}\left(\mathbf{M}_\rho(\mathbf{x},\mathbf{y}) \le 0\right) \;\le\; \underset{(\mathbf{x},\mathbf{y})\sim D}{\Pr}\left(\mathbf{M}_{\rho,\omega}(\mathbf{x},\mathbf{y}) \le 0\right).$$

Therefore, the proof process of Theorem 2 applies if $\underset{(\mathbf{x},\mathbf{y})\sim D}{\mathbf{E}} \mathbf{M}_{\rho,\omega}(\mathbf{x},\mathbf{y}) > 0$.

Note that, even for values of ω for which $\mathcal{C}(\mathbf{M}_{\rho,\omega})$ does not give rise to a valid upper bound of $\mathbf{R}_D(\mathbf{B}_\rho)$, it is still interesting to calculate it as there are some values that can be very good estimators of $\mathbf{R}_D(\mathbf{B}_\rho)$ simultaneously for many different values of ρ. If so, one can capture the behavior of the ensemble method this way. We provide some evidence about this in Section 4.2.

We now theoretically and empirically illustrate these results by addressing the multiclass classification issue from our previous general $\mathcal{C}$-bound perspective.

4 Specializations of the $\mathcal{C}$-bound to Multiclass Prediction

4.1 From Multiclass Margins to $\mathcal{C}$-bounds

The input space at hand still is $\mathbf{X}$, but the output space $\mathbf{Y} = \{1, \ldots, Q\}$ is now made of a finite number of classes (or categories) $Q \ge 2$. We define the output feature map $Y(\cdot)$ such that the image of $\mathbf{Y}$ is $\text{Im}\,\mathbf{Y} = \{0, 1\}^Q$. More precisely, the image of a class $\mathbf{y} \in \mathbf{Y}$ under $Y(\cdot)$ is the canonical Q-dimensional vector $(0, \ldots, 1, \ldots 0)^\top$ whose only nonzero entry is a 1 at its $\mathbf{y}$-th position. The set $\boldsymbol{\mathcal{H}}$ is a set of multiclass voters $\mathbf{h}$ from $\mathbf{X}$ to $\text{conv}(\text{Im}\,\mathbf{Y})$. We recall that given a prior distribution π over $\boldsymbol{\mathcal{H}}$ and an *i.i.d.* m-sample S (drawn from D), the goal of the PAC-Bayesian theory is to estimate the prediction ability of the ρ-weighted majority vote $\mathbf{B}_\rho(\cdot)$ of Equation (2). In this multiclass setting, since for each $\mathbf{c} \in \mathbf{Y}$ only the $\mathbf{c}$-th coordinate of $Y(\mathbf{c})$ equals to 1, the definitions of the majority vote classifier and the margin can

respectively be rewritten as:

$$\mathbf{B}_\rho(\mathbf{x}) \;=\; \operatorname*{argmax}_{\mathbf{c}\in\mathbf{Y}} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}_\mathbf{c}(\mathbf{x})\,,$$
$$\textit{textand} \quad \mathbf{M}_\rho(\mathbf{x},\mathbf{y}) \;=\; \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}_\mathbf{y}(\mathbf{x}) - \max_{\mathbf{c}\in\mathbf{Y},\mathbf{c}\neq\mathbf{y}} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}_\mathbf{c}(\mathbf{x})\,,$$

where $\mathbf{h}_\mathbf{c}(\mathbf{x})$ is the $\mathbf{c}$-th coordinate of $\mathbf{h}(\mathbf{x})$. The following theorem relates $\mathbf{R}_D(\mathbf{B}_\rho)$ and the ω-margin associated to the distribution ρ over $\boldsymbol{\mathcal{H}}$.

Theorem 4. *Let $Q \geq 2$ be the number of classes. For every distribution D over $\mathbf{X}\times\mathbf{Y}$ and for every distribution ρ over a set of multiclass voters $\boldsymbol{\mathcal{H}}$, we have:*

$$\Pr_{(\mathbf{x},\mathbf{y})\sim D}\left(\mathbf{M}_{\rho,\frac{1}{Q}}(\mathbf{x},\mathbf{y}) \leq 0\right) \;\leq\; \mathbf{R}_D(\mathbf{B}_\rho) \;\leq\; \Pr_{(\mathbf{x},\mathbf{y})\sim D}\left(\mathbf{M}_{\rho,\frac{1}{2}}(\mathbf{x},\mathbf{y}) \leq 0\right).$$

Proof. First, let us prove the left-hand side inequality.

$$\begin{aligned}
\mathbf{R}_D(\mathbf{B}_\rho) &= \Pr_{(\mathbf{x},\mathbf{y})\sim D}\left(\mathbf{M}_\rho(\mathbf{x},\mathbf{y}) \leq 0\right)\\
&= \Pr_{(\mathbf{x},\mathbf{y})\sim D}\left(\mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}_\mathbf{y}(\mathbf{x}) \leq \max_{\mathbf{c}\in\mathbf{Y},\mathbf{c}\neq\mathbf{y}} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}_\mathbf{c}(\mathbf{x})\right)\\
&\geq \Pr_{(\mathbf{x},\mathbf{y})\sim D}\left(\mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}_\mathbf{y}(\mathbf{x}) \leq \frac{1}{Q-1}\sum_{\mathbf{c}=1,\mathbf{c}\neq\mathbf{y}}^{Q} \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}_\mathbf{c}(\mathbf{x})\right)\\
&= \Pr_{(\mathbf{x},\mathbf{y})\sim D}\left(\mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}_\mathbf{y}(\mathbf{x}) \leq \frac{1}{Q-1}\left[1 - \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}_\mathbf{y}(\mathbf{x})\right]\right)\\
&= \Pr_{(\mathbf{x},\mathbf{y})\sim D}\left(\mathop{\mathbf{E}}_{\mathbf{h}\sim\rho} \mathbf{h}_\mathbf{y}(\mathbf{x}) - \frac{1}{Q} \leq 0\right)\\
&= \Pr_{(\mathbf{x},\mathbf{y})\sim D}\left(\mathbf{M}_{\rho,\frac{1}{Q}}(\mathbf{x},\mathbf{y}) \leq 0\right).
\end{aligned}$$

The right-hand side inequality is verified by observing that $\mathbf{B}_\rho(\cdot)$ necessarily makes a correct prediction if the weight $\mathbf{E}_{\mathbf{h}\sim\rho}\,\mathbf{h}_\mathbf{y}(\mathbf{x})$ given to the correct $\mathbf{y}$ is higher than $\frac{1}{2}$. □

Consequently, as illustrated in Figure 1, the ω-margin of the points that lie between the $\frac{1}{Q}$-margin and the $\frac{1}{2}$-margin can be negative or positive according to ω. We thus have the following bound.

Corollary 1 (ω-margin multiclass $\mathcal{C}$-bound)**.** *For every probability distribution ρ on a set of multiclass voters $\boldsymbol{\mathcal{H}}$, and for every distribution D on $\mathbf{X}\times\mathbf{Y}$, if $\mathbf{E}_{(\mathbf{x},\mathbf{y})\sim D}\,\mathbf{M}_{\rho,\frac{1}{2}}(\mathbf{x},\mathbf{y}) > 0$, then we have:*

$$\mathbf{R}_D(\mathbf{B}_\rho) \;\leq\; \mathcal{C}(\mathbf{M}_{\rho,\frac{1}{2}}) \;=\; 1 - \frac{\left(\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim D} \mathbf{M}_{\rho,\frac{1}{2}}(\mathbf{x},\mathbf{y})\right)^2}{\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D}\left(\mathbf{M}_{\rho,\frac{1}{2}}(\mathbf{x},\mathbf{y})\right)^2},$$

where $\mathcal{C}(\cdot)$ is the function involved in the ω-margin-based $\mathcal{C}$-bound (Equation (5)).

The region of indecision when $\omega \in [\frac{1}{Q}, \frac{1}{2}]$ implies there is possibly some value of ω to be chosen carefully to provide a good estimator of the true margin. If this is so, we can then consider to make use of $\mathcal{C}(\mathbf{M}_{\rho,\omega})$ for that particular value of ω to improve the analysis of the majority vote's behavior. Obviously, in such a situation, $\mathcal{C}(\mathbf{M}_{\rho,\omega})$ is no longer a bound on $\mathbf{R}_D(\mathbf{B}_\rho)$. However, due to the linearity of the ω-margin, this could open the way to a generalization of the algorithm MinCq (Laviolette et al., 2011) to the multiclass setting.

4.2 Experimental Evaluation of the Bounds

The binary $\mathcal{C}$-bound is known to be well-suited to characterize the behavior of the risk of the ρ-weighted majority vote, as their respective values are correlated (Lacasse et al., 2007). We extend this analysis by empirically evaluating the behavior of the multiclass $\mathcal{C}$-bounds introduced above on natural data. We generate multiclass ρ-weighted majority votes by running a multiclass version of AdaBoost (Freund &

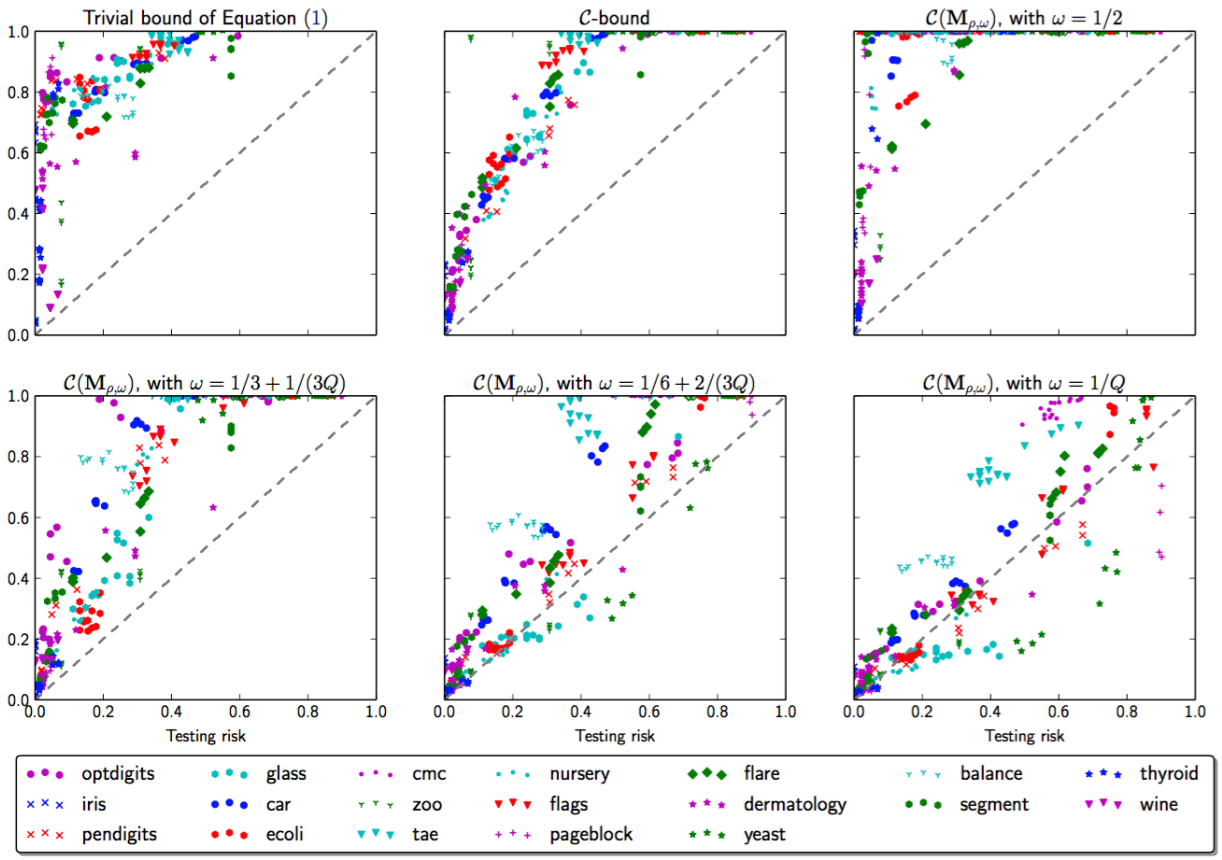

Figure 2: Comparison of the *true* risk of the ρ-weighted majority with: the trivial bound of Equation (1), the $\mathcal{C}$-bound, and $\mathcal{C}(\mathbf{M}_{\rho,\omega})$ for various values of ω. All the values was calculated on a testing set disjoint from the one used to learn ρ.

Schapire, 1997)—known as AdaBoost-SAMME[1] (Zhu et al., 2009)—on multiclass datasets from the UCI dataset repository (Blake & Merz, 1998). We split each dataset in two halves: a training set S and a testing set T. We train the algorithm on S, using 100, 250, 500 and 1,000 decision trees of depth 2, 3, 4 and 5 as base voters, for a total of sixteen majority votes per dataset. The reported values are all computed on the testing set. Figure 2 shows how the values of different upper bounds relate with the risk of the majority vote, and how the choice of ω for various values of $\mathcal{C}(\mathbf{M}_{\rho,\omega})$ (it is not an upper bound for $\omega < \frac{1}{2}$) affects the correlation with the risk. We finalize this correlation study by reporting, in Table 1, the Pearson product-moment correlation coefficients for all computed values.

As pointed out in the paper, we notice from Figure 2 and Table 1 that for some values ω, the values of $\mathcal{C}(\mathbf{M}_{\rho,\omega})$ are very correlated with the risk of the majority vote. Unfortunately, the only one that is an upper bound of the latter ($\omega = \frac{1}{2}$) does not show the same predictive power. Thus, these results also gives some empirical evidence that a wise choice of ω can improve the correlation between the $\mathcal{C}$-bound based on the ω-margin and the risk of the vote.

These experiments confirm the usefulness of the $\mathcal{C}$-bounds based on a notion of margin to upper-bound the true risk of the ρ-weighted majority vote. Taking into account the first and second statistical moments of such margins seems effectively very informative. This property is interesting in an algorithmic point of view: one could derive a multiclass optimization algorithm generalizing the algorithm MinCq (Laviolette et al., 2011) by minimizing $\mathcal{C}(\mathbf{M}_{\rho,\omega})$ where ω could be a hyperparameter to tune by cross-validation.

5 Specializations of the $\mathcal{C}$-bound to Multilabel Prediction

In this section, we instantiate the general $\mathcal{C}$-bound approach to multilabel classification. We stand in the following setting, where the space of possible labels is $\{1, \ldots, Q\}$ with a finite number of classes $Q \geq 2$,

[1] We use of the implementation provided in the Scikit-Learn Python library (Pedregosa et al., 2011).

Table 1: Pearson correlations with the testing risk of the majority vote.

Quantity (evaluated on set T)	Pearson correlation with $\mathbf{R}_T(\mathbf{B}_\rho)$
Trivial bound of Equation (1)	0.7525
$\mathcal{C}(\mathbf{M}_\rho)$, the Multiclass (the $\mathcal{C}$-bound of Theorem 2)	0.8791
$\mathcal{C}(\mathbf{M}_{\rho,\omega})$ with $\omega = 1/2$ (The bound of Corollary 1)	0.5688
$\mathcal{C}(\mathbf{M}_{\rho,\omega})$ with $\omega = 1/3 + 1/(3Q)$ (not a bound)	0.8838
$\mathcal{C}(\mathbf{M}_{\rho,\omega})$ with $\omega = 1/6 + 2/(3Q)$ (not a bound)	**0.9090**
$\mathcal{C}(\mathbf{M}_{\rho,\omega})$ with $\omega = 1/Q$ (not a bound)	0.8741

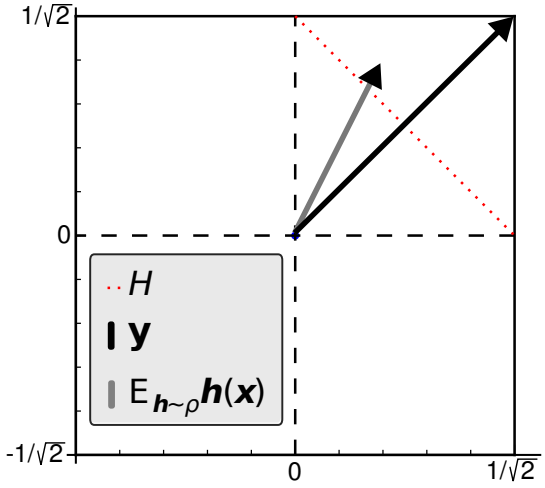

Figure 3: Graphical representation of the $\frac{Q-1}{Q}$-margin and the vote applied on an example $(\mathbf{x}, \mathbf{y})$ for multilabel classification when $Q = 2$ and the true $\mathbf{y}$ is $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^\top$. The angles of the cube corresponds to the different multilabels, that are: $Y(\mathbf{Y}) = \{(\frac{-1}{\sqrt{2}}, \frac{-1}{\sqrt{2}})^\top, (\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}})^\top, (\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^\top, (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^\top\}$. Each line represents the decision boundary of a margin: the hyperplane where lies each example with a margin equals to 0. A vote correctly classifies an example if it lies on the same side of the hyperplane than the correct class.

but we consider the *multilabel* output space $\mathbf{Y} = \{0, 1\}^Q$ that contains vectors $\mathbf{y} = (y_1, \ldots, y_Q)^\top$. In other words we consider multiple binary labels. Given an example $(\mathbf{x}, \mathbf{y}) \in \mathbf{X} \times \mathbf{Y}$, the output vector $\mathbf{y}$ is then defined as follows:

$$\forall j \in \{1, \ldots, Q\}, \quad y_j = \begin{cases} 1 \text{ if } \mathbf{x} \text{ is labeled with } j \\ 0 \text{ otherwise.} \end{cases}$$

In this specific case, we define the output feature map $Y(\cdot)$ such that the image of $\mathbf{Y}$ is $\operatorname{Im} \mathbf{Y} = \left\{\frac{-1}{\sqrt{Q}}, \frac{1}{\sqrt{Q}}\right\}^Q$, and:

$$\forall j \in \{1, \ldots, Q\}, \quad Y_j(\mathbf{y}) = \begin{cases} \frac{1}{\sqrt{Q}} \text{ if } y_j = 1 \ (\mathbf{x} \text{ is labeled with } j) \\ \frac{-1}{\sqrt{Q}} \text{ otherwise,} \end{cases}$$

where $Y_j(\mathbf{y})$ is the j-th coordinate of $Y(\mathbf{y})$. According to this definition, we have that: $\forall \mathbf{y} \in \mathbf{Y}$, $\|Y(\mathbf{y})\| = 1$. The set $\boldsymbol{\mathcal{H}}$ is made of *multilabel voters* $\mathbf{h} : \mathbf{X} \to \operatorname{conv}(\operatorname{Im} \mathbf{Y})$. In the light of the feature output map $Y(\cdot)$, the definition of the majority vote classifier and the margin can respectively be rewritten as:

$$\mathbf{B}_\rho(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{c} \in \mathbf{Y}} \sum_{j=1}^{Q} \mathop{\mathbf{E}}_{\mathbf{h} \sim \rho} \mathbf{h}_j(\mathbf{x}) Y_j(\mathbf{y}) \,,$$

$$\text{and} \quad \mathbf{M}_\rho(\mathbf{x}, \mathbf{y}) = \sum_{j=1}^{Q} \mathop{\mathbf{E}}_{\mathbf{h} \sim \rho} \mathbf{h}_j(\mathbf{x}) Y_j(\mathbf{y}) - \max_{\substack{\mathbf{c} \in \mathbf{Y} \\ \mathbf{c} \neq \mathbf{y}}} \sum_{j=1}^{Q} \mathop{\mathbf{E}}_{\mathbf{h} \sim \rho} \mathbf{h}_j(\mathbf{x}) Y_j(\mathbf{c}) \,,$$

where $\mathbf{h}_j$ is the j-th coordinate of $\mathbf{h}(\mathbf{x})$.

The next theorem relates the risk of $\mathbf{B}_\rho(\cdot)$ and the ω-margin associated to ρ.

Theorem 5. *Let $Q \geq 2$ be the number of labels. For every distribution D over $\mathbf{X} \times \mathbf{Y}$ and for every distribution ρ over a set of multilabel voters $\mathcal{H}$, we have:*

$$\mathbf{R}_D(\mathbf{B}_\rho) \leq \Pr_{(\mathbf{x},\mathbf{y})\sim D} \left(\mathbf{M}_{\rho,\frac{Q-1}{Q}}(\mathbf{x},\mathbf{y}) \leq 0 \right) .$$

Proof. We have to show:

$$\Pr_{(\mathbf{x},\mathbf{y})\sim D} \left(\mathbf{M}_{\rho}(\mathbf{x},\mathbf{y}) \leq 0 \right) \leq \Pr_{(\mathbf{x},\mathbf{y})\sim D} \left(\mathbf{M}_{\rho,\frac{Q-1}{Q}}(\mathbf{x},\mathbf{y}) \leq 0 \right) .$$

To do so we will prove that:

$$\mathbf{M}_{\rho,\frac{Q-1}{Q}}(\mathbf{x},\mathbf{y}) > 0 \Longrightarrow \mathbf{M}_{\rho}(\mathbf{x},\mathbf{y}) > 0 .$$

Recall that $\mathrm{conv}(\mathrm{Im}\,\mathbf{Y})$ is a hypercube whose vertices are exactly the $Y(\mathbf{c})$'s with $\mathbf{c} \in \mathbf{Y}$. Given a vertex $Y(\mathbf{y})$, denote $H_\mathbf{y}$ the hyperplane who passes through all the points $Y^{(j)}(\mathbf{y})$, where $Y^{(j)}(\mathbf{y})$ is the point of the hypercube that has exactly the same coordinates as $Y(\mathbf{y})$, except the j^{th} that has been put to 0.

Now, consider the region $R_\mathbf{y}$ of the hypercube $\mathrm{conv}(\mathrm{Im}\,\mathbf{Y})$ that consists of all the points that correspond to $\mathbf{M}_{\rho,\frac{Q-1}{Q}}(\mathbf{x},\mathbf{y}) > 0$, that is, the points that are on the same side of hyperplane $H_\mathbf{y}$ than $Y(\mathbf{y})$. Clearly, for any $Q \geq 2$, $Y(\mathbf{y})$ is strictly closer to the point $\mathbf{E}_{\mathbf{h}\sim\rho}\,\mathbf{h}(\mathbf{x})$ than any other $Y(\mathbf{c})$'s if the vector $\mathbf{E}_{\mathbf{h}\sim\rho}\,\mathbf{h}(\mathbf{x})$ lies in $R_\mathbf{y}$. This in turn implies that the margin $\mathbf{M}_\rho(\mathbf{x},\mathbf{y})$ is strictly positive. Figure 3 shows an example with $Q = 2$ and $Y(\mathbf{y}) = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$, where $H_\mathbf{y}$ is represented by a red dotted line, and $R_\mathbf{y}$ is the region delimited by the top-right corner and $H_\mathbf{y}$.

To finish the proof, we have to show that $R_\mathbf{y}$ is exactly the region where $\mathbf{M}_{\rho,\frac{Q-1}{Q}}(\mathbf{x},\mathbf{y}) > 0$. Equivalently, we must show that the intersection of $H_\mathbf{y}$ and the hypercube $\mathrm{conv}(\mathrm{Im}\,\mathbf{Y})$ is exactly the points for which $\mathbf{M}_{\rho,\frac{Q-1}{Q}}(\mathbf{x},\mathbf{y}) = 0$, *i.e.,* the vectors $\mathbf{E}_{\mathbf{h}\sim\rho}\,\mathbf{h}(\mathbf{x})$ for which $\langle \mathbf{E}_{\mathbf{h}\sim\rho}\,\mathbf{h}(\mathbf{x})\,,\,\mathbf{y}\rangle - \frac{Q-1}{Q} = 0$. We know from basic linear algebra that the points P that lie on hyperplane $H_\mathbf{y}$ must satisfy the following equation: $(P - P_0)\cdot N = 0$, where N is the normal of the hyperplane and P_0 is any point in P. It is easy to see that $Y(\mathbf{y})$ is the normal of H and that we can take $P_0 = Y^{(1)}(\mathbf{y})$. Hence, the equation becomes $(P - Y^{(1)}(\mathbf{y}))\cdot Y(\mathbf{y}) = 0$.

Since all coordinates of $Y(\mathbf{y})$ are either $\frac{1}{\sqrt{Q}}$ or $\frac{-1}{\sqrt{Q}}$, and all coordinates of $Y^{(1)}(\mathbf{y})$ are the same as the ones of $Y(\mathbf{y})$ except the first one being 0 in $Y^{(1)}(\mathbf{y})$, we have that $Y^{(1)}(\mathbf{y})\cdot Y(\mathbf{y}) = \frac{Q-1}{Q}$. The result then follows from

$$(P - Y^{(1)}(\mathbf{y}))\cdot Y(\mathbf{y}) \;=\; P\cdot Y(\mathbf{y}) - Y^{(1)}(\mathbf{y})\cdot Y(\mathbf{y}) \;=\; \langle P\,,\,Y(\mathbf{y})\rangle - \frac{Q-1}{Q} .$$

□

Finally, according to the same arguments as in Corollary 1, we have:

Corollary 2 (ω-margin multilabel $\mathcal{C}$-bound)**.** *For every probability distribution ρ on a set of multilabel voters $\mathcal{H}$, for every distribution D on $\mathbf{X}\times\mathbf{Y}$, if $\mathbf{E}_{(\mathbf{x},\mathbf{y})\sim D}\,\mathbf{M}_{\rho,\frac{Q-1}{Q}}(\mathbf{x},\mathbf{y}) > 0$, we have:*

$$\mathbf{R}_D(\mathbf{B}_\rho) \leq\; \mathcal{C}(\mathbf{M}_{\rho,\frac{Q-1}{Q}}) \;=\; 1 - \frac{\left(\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim D} \mathbf{M}_{\rho,\frac{Q-1}{Q}}(\mathbf{x},\mathbf{y})\right)^2}{\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} \left(\mathbf{M}_{\rho,\frac{Q-1}{Q}}(\mathbf{x},\mathbf{y})\right)^2} .$$

6 Conclusion

In PAC-Bayesian binary classification, it is well-known that the $\mathcal{C}$-bound offers a tight bound over the risk of the ρ-weighted majority vote by taking into account the first two statistical moments of its margin. Moreover, from a practical standpoint, minimizing the $\mathcal{C}$-bound leads to a well-performing algorithm called MinCq (Laviolette et al., 2011). This paper fills the gap between this binary classification theory and more complex tasks by generalizing the $\mathcal{C}$-bound for majority vote over complex output voters, and by proposing a new surrogate of the margin easier to manipulate. Note that, as future work, we would like to study how tuning this surrogate margin would result in a precise estimation of the risk. In order to justify the empirical estimation of the $\mathcal{C}$-bound from a sample, we provide a PAC-Bayesian generalization bound. Moreover, we show how to specialize our result to multiclass and multilabel classification. Concretely, we

think that the theoretical $\mathcal{C}$-bounds provided here are a first step towards developing ensemble methods to learn ρ-weighted majority vote for complex outputs through the minimization of a $\mathcal{C}$-bound, or of a surrogate of it. A first solution for deriving such a method could be to study the general weak learning conditions necessary and sufficient to define an ensemble of structured output voters, as done by Mukherjee and Shapire (Mukherjee & Schapire, 2013) for multiclass boosting.

Appendix

A The Bounds Requiered to Prove Theorem 3

Theorem 6. *For any distribution D on $\mathbf{X}\times\mathbf{Y}$, for any set $\boldsymbol{\mathcal{H}}$ of voters from $\mathbf{X}$ to $\mathrm{conv}(\mathrm{Im}\,\mathbf{Y})$, for any prior distribution π on $\boldsymbol{\mathcal{H}}$ and any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the choice of the m-sample $S\sim(D)^m$, for every posterior distribution ρ over $\boldsymbol{\mathcal{H}}$ we have :*

$$\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\in D}\mathbf{M}_\rho(\mathbf{x},\mathbf{y}) \;\geq\; \frac{1}{m}\sum_{(\mathbf{x},\mathbf{y})\in S}\mathbf{M}_\rho(\mathbf{x},\mathbf{y}) - \sqrt{\frac{2B}{m}\left[\mathrm{KL}(\rho\|\pi)+\ln\frac{2\sqrt{m}}{\delta}\right]},$$

where $B\in(0,2]$ bounds the absolute value of the margin $|\mathbf{M}_\rho(\mathbf{x},\mathbf{y})|$ for all $(\mathbf{x},\mathbf{y})$, and $\mathrm{KL}(\rho\|\pi)=\mathbf{E}_{\mathbf{h}\sim\rho}\ln\frac{\rho(\mathbf{h})}{\pi(\mathbf{h})}$ is the Kullback-Leibler divergence between ρ and π.

Proof. The following proof shows how to obtain the lower bound on the first moment of $\mathbf{M}_\rho(\mathbf{x},\mathbf{y})$, and uses the same notions as the classical PAC-Bayesian proofs.[2]

Given a distribution D' on $\mathbf{X}\times\mathbf{Y}$, for any distribution ρ' over $\boldsymbol{\mathcal{H}}$, we can rewrite $\mathbf{E}_{(\mathbf{x},\mathbf{y})\sim D'}\mathbf{M}_{\rho'}(\mathbf{x},\mathbf{y})$ as an expectation over ρ'. We denote $\mathbf{M}_{\mathbf{h}}^{D'}$ the random variable such that $\mathbf{E}_{\mathbf{h}\sim\rho'}\mathbf{M}_{\mathbf{h}}^{D'}=\mathbf{E}_{(\mathbf{x},\mathbf{y})\sim D'}\mathbf{M}_{\rho'}(\mathbf{x},\mathbf{y})$.

First, we have that $\mathbf{E}_{\mathbf{h}\sim\pi}\exp\left[\frac{m}{2B}\left(\mathbf{M}_{\mathbf{h}}^S-\mathbf{M}_{\mathbf{h}}^D\right)^2\right]$ is a non-negative random variable. Applying Markov's inequality yields that with probability at least $1-\delta$ over the choice of $S\sim(D)^m$, we have:

$$\mathop{\mathbf{E}}_{\mathbf{h}\sim\pi}\exp\left[\frac{m}{2B}\left(\mathbf{M}_{\mathbf{h}}^S-\mathbf{M}_{\mathbf{h}}^D\right)^2\right] \;\leq\; \frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{\mathbf{h}\sim\pi}\exp\left[\frac{m}{2B}\left(\mathbf{M}_{\mathbf{h}}^S-\mathbf{M}_{\mathbf{h}}^D\right)^2\right]. \tag{6}$$

We upper-bound the right-hand side of the inequality:

$$\begin{aligned}
\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{\mathbf{h}\sim\pi}\exp\left[\frac{m}{2B}\left(\mathbf{M}_{\mathbf{h}}^S-\mathbf{M}_{\mathbf{h}}^D\right)^2\right] &= \mathop{\mathbf{E}}_{\mathbf{h}\sim\pi}\mathop{\mathbf{E}}_{S\sim D^m}\exp\left[\frac{m}{2B}\left(\mathbf{M}_{\mathbf{h}}^S-\mathbf{M}_{\mathbf{h}}^D\right)^2\right] &(7)\\
&= \mathop{\mathbf{E}}_{\mathbf{h}\sim\pi}\mathop{\mathbf{E}}_{S\sim D^m}\exp\left[m\,2\left(\frac{1}{2}\left(1-\frac{1}{B}\mathbf{M}_{\mathbf{h}}^S\right)-\frac{1}{2}\left(1-\frac{1}{B}\mathbf{M}_{\mathbf{h}}^D\right)\right)^2\right]\\
&\leq \mathop{\mathbf{E}}_{\mathbf{h}\sim\pi}\mathop{\mathbf{E}}_{S\sim D^m}\exp\left[m\,\mathrm{kl}\left(\frac{1}{2}\left(1-\frac{\mathbf{M}_{\mathbf{h}}^S}{B}\right)\middle\|\frac{1}{2}\left(1-\frac{\mathbf{M}_{\mathbf{h}}^D}{B}\right)\right)\right] &(8)\\
&\leq \mathop{\mathbf{E}}_{\mathbf{h}\sim\pi}2\sqrt{m} \;=\; 2\sqrt{m}\,. &(9)
\end{aligned}$$

Line (7) comes from the fact that the distribution π is defined a priori. Since B is an upper bound of the possible absolute values of the margin, both $\frac{1}{2}\left(1-\frac{\mathbf{M}_{\mathbf{h}}^S}{B}\right)$ and $\frac{1}{2}\left(1-\frac{\mathbf{M}_{\mathbf{h}}^D}{B}\right)$ are between 0 and 1. Thus Line (8) is an application of Pinsker's inequality $2(q-p)^2\leq\mathrm{kl}(q\|p)=q\ln\frac{q}{p}+(1-q)\ln\frac{1-q}{1-p}$. Finally, Line (9) is an application of Maurer (2004) (Theorem 5.).

By applying this upper bound in Inequality (6) and by taking the logarithm on each side, with probability at least $1-\delta$ over the choice of $S\sim D^m$, we have:

$$\ln\left(\mathop{\mathbf{E}}_{\mathbf{h}\sim\pi}\exp\left[\frac{m}{2B}\left(\mathbf{M}_{\mathbf{h}}^S-\mathbf{M}_{\mathbf{h}}^D\right)^2\right]\right) \;\leq\; \ln\left(\frac{2\sqrt{m}}{\delta}\right).$$

Now, by applying the change of measure inequality proposed by Seldin & Tishby (2010) (Lemma 4.) with $\phi(\mathbf{h})=\frac{m}{2B}\left(\mathbf{M}_{\mathbf{h}}^S-\mathbf{M}_{\mathbf{h}}^D\right)^2$, and by using Jensen's inequality exploiting the convexity of $\phi(\mathbf{h})$, we obtain

[2] The reader can refer to (Germain et al., 2009; Seeger, 2003; Catoni, 2007; McAllester, 2003) for examples of classical PAC-Bayesian analyses.

that for all distributions ρ on $\boldsymbol{\mathcal{H}}$:

$$\ln\left(\mathop{\mathbf{E}}_{\mathbf{h}\sim\pi}\exp\left[\frac{m}{2B}\left(\mathbf{M}^S_{\mathbf{h}}-\mathbf{M}^D_{\mathbf{h}}\right)^2\right]\right) \;\geq\; \mathop{\mathbf{E}}_{\mathbf{h}\sim\rho}\frac{m}{2B}\left(\mathbf{M}^S_{\mathbf{h}}-\mathbf{M}^D_{\mathbf{h}}\right)^2-\mathrm{KL}(\rho\|\pi)$$
$$\geq\; \frac{m}{2B}\left(\mathop{\mathbf{E}}_{\mathbf{h}\sim\rho}\mathbf{M}^S_{\mathbf{h}}-\mathop{\mathbf{E}}_{\mathbf{h}\sim\rho}\mathbf{M}^D_{\mathbf{h}}\right)^2-\mathrm{KL}(\rho\|\pi)\,.$$

From all what precedes, we have that with probability at least $1-\delta$ over the choice of $S\sim(D)^m$, for every posterior distribution ρ on $\boldsymbol{\mathcal{H}}$, we have:

$$\frac{m}{2B}\left(\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim S}\mathbf{M}_\rho(\mathbf{x},\mathbf{y})-\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim D}\mathbf{M}_\rho(\mathbf{x},\mathbf{y})\right)^2-\mathrm{KL}(\rho\|\pi)\;\leq\;\ln\left(\frac{2\sqrt{m}}{\delta}\right).$$

The result follows from algebraic calculations. □

Theorem 7. *For any distribution D on $\mathbf{X}\times\mathbf{Y}$, for any set $\boldsymbol{\mathcal{H}}$ of voters from $\mathbf{X}$ to $\mathrm{conv}(\mathrm{Im}\,\mathbf{Y})$, for any prior distribution π on $\boldsymbol{\mathcal{H}}$ and any $\delta\in(0,1]$, with a probability at least $1-\delta$ over the choice of the m-sample $S\sim(D)^m$, for every posterior distribution ρ over $\boldsymbol{\mathcal{H}}$ we have :*

$$\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\in S}\left(\mathbf{M}_\rho(\mathbf{x},\mathbf{y})\right)^2\;\leq\;\frac{1}{m}\sum_{(\mathbf{x},\mathbf{y})\in S}\left(\mathbf{M}_\rho(\mathbf{x},\mathbf{y})\right)^2+\sqrt{\frac{2B^2}{m}\left[2\mathrm{KL}(\rho||\pi)+\ln\frac{2\sqrt{m}}{\delta}\right]},$$

where $B\in(0,2]$ bounds the absolute value of the margin $|\mathbf{M}_\rho(\mathbf{x},\mathbf{y})|$ for all $(\mathbf{x},\mathbf{y})$, and $\mathrm{KL}(\rho\|\pi)=\mathbf{E}_{\mathbf{h}\sim\rho}\ln\frac{\rho(\mathbf{h})}{\pi(\mathbf{h})}$ is the Kullback-Leibler divergence between ρ and π.

Proof. This proof uses many notions that are usual in classical PAC-Bayesian proofs, but the expectation over single voters is replaced with an expectation over pairs of voters. Given a distribution D' on $\mathbf{X}\times\mathbf{Y}$, for any distribution ρ'^2 over $\boldsymbol{\mathcal{H}}$, we rewrite $\mathbf{E}_{(\mathbf{x},\mathbf{y})\sim D'}(\mathbf{M}_{\rho'}(\mathbf{x},\mathbf{y}))^2$ as an expectation over ρ'^2. Let $\mathbf{M}^{D'}_{\mathbf{h},\mathbf{h}'}$ be the r.v. such that $\mathbf{E}_{(\mathbf{h},\mathbf{h}')\sim\rho'^2}\mathbf{M}^{D'}_{\mathbf{h},\mathbf{h}'}=\mathbf{E}_{(\mathbf{x},\mathbf{y})\sim D'}(\mathbf{M}_{\rho'}(\mathbf{x},\mathbf{y}))^2$. First, we apply the Markov's inequality on the non-negative r.v. $\mathbf{E}_{(\mathbf{h},\mathbf{h}')\sim\pi^2}\exp\left[\frac{m}{2B^2}\left(\mathbf{M}^S_{\mathbf{h},\mathbf{h}'}-\mathbf{M}^D_{\mathbf{h},\mathbf{h}'}\right)^2\right]$. Thus, we have that with probability at least $1-\delta$ over the choice of $S\sim(D)^m$:

$$\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2}\exp\left[\frac{m}{2B^2}\left(\mathbf{M}^S_{\mathbf{h},\mathbf{h}'}-\mathbf{M}^D_{\mathbf{h},\mathbf{h}'}\right)^2\right]\leq\frac{1}{\delta}\mathop{\mathbf{E}}_{S\sim(D)^m}\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2}\exp\left[\frac{m}{2B^2}\left(\mathbf{M}^S_{\mathbf{h},\mathbf{h}'}-\mathbf{M}^D_{\mathbf{h},\mathbf{h}'}\right)^2\right]. \tag{10}$$

Then, we upper-bound the right-hand side of the inequality:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{S\sim D^m}\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2}\exp\left[\frac{m}{2B^2}\left(\mathbf{M}^S_{\mathbf{h},\mathbf{h}'}-\mathbf{M}^D_{\mathbf{h},\mathbf{h}'}\right)^2\right]\\
&\quad=\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2}\mathop{\mathbf{E}}_{S\sim D^m}\exp\left[\frac{m}{2B^2}\left(\mathbf{M}^S_{\mathbf{h},\mathbf{h}'}-\mathbf{M}^D_{\mathbf{h},\mathbf{h}'}\right)^2\right] &(11)\\
&\quad=\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2}\mathop{\mathbf{E}}_{S\sim D^m}\exp\left[m\,2\left(\frac{1}{2}\left(1-\frac{1}{B^2}\mathbf{M}^S_{\mathbf{h},\mathbf{h}'}\right)-\frac{1}{2}\left(1-\frac{1}{B^2}\mathbf{M}^D_{\mathbf{h},\mathbf{h}'}\right)\right)^2\right]\\
&\quad\leq\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2}\mathop{\mathbf{E}}_{S\sim D^m}\exp\left[m\,\mathrm{kl}\left(\frac{1}{2}\left(1-\frac{\mathbf{M}^S_{\mathbf{h},\mathbf{h}'}}{B^2}\right)\middle\|\frac{1}{2}\left(1-\frac{\mathbf{M}^D_{\mathbf{h},\mathbf{h}'}}{B^2}\right)\right)\right] &(12)\\
&\quad\leq\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2}2\sqrt{m}\;=\;2\sqrt{m}\,. &(13)
\end{aligned}$$

Line 11 comes from the fact that the distribution π is defined *a priori*, *i.e.*, before observing S. Since B upper-bounds the absolute value of the margin, both $\frac{1}{2}\left(1-\frac{\mathbf{M}^S_{\mathbf{h},\mathbf{h}'}}{B^2}\right)$ and $\frac{1}{2}\left(1-\frac{\mathbf{M}^D_{\mathbf{h},\mathbf{h}'}}{B^2}\right)$ lie between 0 and 1. Line 12 is then an application of Pinsker's inequality[3]. Finally, Line 13 is an application of Maurer (2004) (Theorem 5), which is stated to be valid for $m\geq 8$, but is also valid for any $m\geq 1$.
By applying this upper bound in Inequality (10) and by taking the logarithm on each side, with probability at least $1-\delta$ over the choice of $S\sim(D)^m$, we have:

$$\ln\left(\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2}\exp\left[\frac{m}{2B^2}\left(\mathbf{M}^S_{\mathbf{h},\mathbf{h}'}-\mathbf{M}^D_{\mathbf{h},\mathbf{h}'}\right)^2\right]\right)\;\leq\;\ln\left(\frac{2\sqrt{m}}{\delta}\right).$$

[3]The Pinkser inequality is: $2(q-p)^2\leq\mathrm{kl}(q\|p)=q\ln\frac{q}{p}+(1-q)\ln\frac{1-q}{1-p}$

Now, we need the change of measure inequality[4] of Lemma 1 (stated below) that has the novelty to use pairs of voters. By applying this lemma with $\phi(\mathbf{h},\mathbf{h}') = \frac{m}{2B^2}\left(\mathbf{M}^S_{\mathbf{h},\mathbf{h}'} - \mathbf{M}^D_{\mathbf{h},\mathbf{h}'}\right)^2$, and by using Jensen's inequality exploiting the convexity of $\phi(\mathbf{h},\mathbf{h}')$, we obtain that for all distributions ρ on $\boldsymbol{\mathcal{H}}$:

$$\ln\left(\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2}\exp\left[\frac{m}{2B^2}\left(\mathbf{M}^S_{\mathbf{h},\mathbf{h}'} - \mathbf{M}^D_{\mathbf{h},\mathbf{h}'}\right)^2\right]\right) \geq \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2}\frac{m}{2B^2}\left(\mathbf{M}^S_{\mathbf{h},\mathbf{h}'} - \mathbf{M}^D_{\mathbf{h},\mathbf{h}'}\right)^2 - 2\mathrm{KL}(\rho\|\pi)$$
$$\geq \frac{m}{2B^2}\left(\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2}\mathbf{M}^S_{\mathbf{h},\mathbf{h}'} - \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2}\mathbf{M}^D_{\mathbf{h},\mathbf{h}'}\right)^2 - 2\mathrm{KL}(\rho\|\pi)\,.$$

From all what precedes, with probability at least $1-\delta$ on the choice of $S\sim(D)^m$, for every posterior distribution ρ on $\boldsymbol{\mathcal{H}}$, we have:

$$\frac{m}{2B^2}\left(\mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim S}(\mathbf{M}_\rho(\mathbf{x},\mathbf{y}))^2 - \mathop{\mathbf{E}}_{(\mathbf{x},\mathbf{y})\sim D}(\mathbf{M}_\rho(\mathbf{x},\mathbf{y}))^2\right)^2 - 2\mathrm{KL}(\rho\|\pi) \;\leq\; \ln\left(\frac{2\sqrt{m}}{\delta}\right).$$

The result follows from algebraic calculations. □

The change of measure used in the previous proof is stated below.

Lemma 1 (Change of measure inequality for pairs of voters)**.** *For any set of voters $\boldsymbol{\mathcal{H}}$, for any distributions π and ρ on $\boldsymbol{\mathcal{H}}$, and for any measurable function $\phi : \boldsymbol{\mathcal{H}}\times\boldsymbol{\mathcal{H}} \to \mathbb{R}$, we have:*

$$\ln\left(\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2}\exp\left[\phi(\mathbf{h},\mathbf{h}')\right]\right) \;\geq\; \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2}\phi(\mathbf{h},\mathbf{h}') - 2\mathrm{KL}(\pi\|\rho)\,.$$

Proof. The proof is very similar to the one of Seldin & Tishby (2010) (Lemma 4.), but is defined using pairs of voters. The first inequality below is given by using Jensen's inequality on the concave function $\ln(\cdot)$.

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2}\phi(\mathbf{h},\mathbf{h}') &= \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2}\ln\left(\mathrm{e}^{\phi(\mathbf{h},\mathbf{h}')}\right) = \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2}\ln\left(\mathrm{e}^{\phi(\mathbf{h},\mathbf{h}')}\,\frac{\rho^2(\mathbf{h},\mathbf{h}')}{\pi^2(\mathbf{h},\mathbf{h}')}\,\frac{\pi^2(\mathbf{h},\mathbf{h}')}{\rho^2(\mathbf{h},\mathbf{h}')}\right)\\
&= \mathrm{KL}(\rho^2\|\pi^2) + \mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2}\ln\left(\mathrm{e}^{\phi(\mathbf{h},\mathbf{h}')}\,\frac{\pi^2(\mathbf{h},\mathbf{h}')}{\rho^2(\mathbf{h},\mathbf{h}')}\right)\\
&\leq \mathrm{KL}(\rho^2\|\pi^2) + \ln\left(\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\rho^2}\mathrm{e}^{\phi(\mathbf{h},\mathbf{h}')}\,\frac{\pi^2(\mathbf{h},\mathbf{h}')}{\rho^2(\mathbf{h},\mathbf{h}')}\right)\\
&\leq \mathrm{KL}(\rho^2\|\pi^2) + \ln\left(\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2}\mathrm{e}^{\phi(\mathbf{h},\mathbf{h}')}\right)\\
&= 2\mathrm{KL}(\rho\|\pi) + \ln\left(\mathop{\mathbf{E}}_{(\mathbf{h},\mathbf{h}')\sim\pi^2}\mathrm{e}^{\phi(\mathbf{h},\mathbf{h}')}\right).
\end{aligned}$$

Note that the last inequality becomes an equality if ρ and π share the same support. The last equality comes from the definition of the KL-divergence, and from the fact that $\pi^2(\mathbf{h},\mathbf{h}') = \pi(\mathbf{h})\pi(\mathbf{h}')$ and $\rho^2(\mathbf{h},\mathbf{h}') = \rho(\mathbf{h})\rho(\mathbf{h}')$. □

References

Allwein, E.L., Schapire, R.E., and Singer, Y. Reducing multiclass to binary: A unifying approach for margin classifiers. *JMLR*, 1:113–141, 2001.

Blake, C.L. and Merz, C.J. *UCI Repository of machine learning databases.* Dpt. of Information & Computer Science, Univ. of California, `archive.ics.uci.edu/ml`, 1998.

Blanchard, G. Different paradigms for choosing sequential reweighting algorithms. *Neural Comput.*, 16 (4):811–836, 2004.

Boser, B.E., Guyon, I.M., and Vapnik, V.N. A training algorithm for optimal margin classifiers. In *COLT*, pp. 144–152, 1992.

Breiman, L. Bagging predictors. *Mach. Learn.*, 24(2):123–140, 1996.

[4] The change of measure is an important step in most PAC-Bayesian proofs (Seldin & Tishby, 2010).

Breiman, L. Random Forests. *Mach. Learn.*, 45(1):5–32, 2001.

Catoni, O. *PAC-Bayesian supervised classification: the thermodynamics of statistical learning*, volume 56. IMS Lecture Notes Monogr. Ser., 2007. ISBN 9780940600720.

Cortes, C. and Vapnik, V. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995.

Cortes, Corinna, Kuznetsov, Vitaly, and Mohri, Mehryar. Ensemble methods for structured prediction. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1134–1142, 2014.

Dietterich, T.G. Ensemble methods in machine learning. In *MCS*, pp. 1–15. Springer, 2000.

Dietterich, T.G. and Bakiri, G. Solving multiclass learning problems via error-correcting output codes. *J. Artif. Intell. Res.*, 2(263):286, 1995.

Domingos, P. Bayesian averaging of classifiers and the overfitting problem. In *ICML*, pp. 223–230, 2000.

Freund, Y. and Schapire, R.E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55:119–139, 1997.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. *Bayesian data analysis*. Chapman & Hall/CRC, 2004. ISBN 9781584883883.

Germain, P., Lacasse, A., Laviolette, F., and Marchand, M. PAC-Bayesian learning of linear classifiers. In *ICML*, pp. 353–360, 2009.

Haussler, D., Kearns, M., and Schapire, R.E. Bounds on the sample complexity of bayesian learning using information theory and the VC dimension. *Mach. Learn.*, 14(1):83–113, 1994.

Kuznetsov, V., Mohri, M., and Syed, U. Multi-class deep boosting. In *NIPS*, pp. 2501–2509, 2014.

Lacasse, A., Laviolette, F., Marchand, M., Germain, P., and Usunier, N. PAC-Bayes bounds for the risk of the majority vote and the variance of the Gibbs classifier. In *NIPS*, 2007.

Langford, J. and Shawe-Taylor, J. PAC-Bayes & margins. In *NIPS*, pp. 423–430, 2002.

Laviolette, F., Marchand, M., and Roy, J.-F. From PAC-Bayes bounds to quadratic programs for majority votes. In *ICML*, 2011.

Maurer, Andreas. A note on the PAC-Bayesian theorem. *CoRR*, cs.LG/0411099, 2004.

McAllester, D. Simplified PAC-Bayesian margin bounds. In *COLT*, pp. 203–215, 2003.

McAllester, D.A. Some PAC-Bayesian theorems. *Mach. Learn.*, 37:355–363, 1999.

McAllester, David. Generalization bounds and consistency for structured labeling. 2009.

Morvant, E., Koço, S., and Ralaivola, L. PAC-Bayesian generalization bound on confusion matrix for multi-class classification. In *ICML*, 2012.

Morvant, E., Habrard, A., and Ayache, S. Majority vote of diverse classifiers for late fusion. In *IAPR S+SSPR*, 2014.

Mroueh, Y., Poggio, T., Rosasco, L., and Slotine, J.-J. Multiclass learning with simplex coding. In *NIPS*, pp. 2789–2797, 2012.

Mukherjee, I. and Schapire, R.E. A theory of multiclass boosting. *JMLR*, 14(1):437–497, 2013.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830, 2011.

Re, M. and Valentini, G. Ensemble methods: a review. 2012.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. Classifier chains for multi-label classification. *Mach. Learn.*, 85(3):333–359, 2011.

Schapire, R.E. and Singer, Y. Improved boosting algorithms using confidence-rated predictions. In *Mach. Learn.*, pp. 80–91, 1999.

Seeger, M. PAC-Bayesian generalisation error bounds for gaussian process classification. *JMLR*, 3:233–269, 2003.

Seldin, Y. and Tishby, N. PAC-Bayesian analysis of co-clustering and beyond. *JMLR*, 11:3595–3646, 2010.

Sun, S. A survey of multi-view machine learning. *Neural Comput. Appl.*, 23(7-8):2031–2038, 2013.

Tsoumakas, G. and Vlahavas, I.P. Random k-labelsets: An ensemble method for multilabel classification. In *ECML*, pp. 406–417, 2007.

Zhang, Y. and Schneider, J. Maximum margin output coding. In *ICML*, pp. 1575–1582, 2012.

Zhu, J., Zou, H., Rosset, S., and Hastie, T. Multi-class AdaBoost. *Stat. Interface*, 2(3):349–360, 2009.